DOCUMENT RESUME

ED 481 063 TM 035 241

AUTHOR Zwick, Rebecca; Thayer, Dorothy T.

TITLE An Empirical Bayes Enhancement of Mantel-Haenszel DIF Analysis for Computer-Adaptive Tests. LSAC Research Report Series.
INSTITUTION Law School Admission Council, Newtown, PA.
REPORT NO LSAC-RR-98-15
PUB DATE 2003-08-00
NOTE 31p.
PUB TYPE Reports - Research (143)
EDRS PRICE EDRS Price MF01/PC02 Plus Postage.
DESCRIPTORS *Adaptive Testing; *Bayesian Statistics; College Entrance Examinations; *Computer Assisted Testing; *Item Bias; Simulation
IDENTIFIERS *Law School Admission Test; *Mantel Haenszel Procedure

ABSTRACT

This study investigated the applicability to computerized adaptive testing (CAT) data of a differential item functioning (DIF) analysis that involves an empirical Bayes (EB) enhancement of the popular Mantel Haenszel (MH) DIF analysis method. The computerized Law School Admission Test (LSAT) assumed for this study was similar to that currently being evaluated for a potential computerized LSAT. In this case, rather than being presented with a single item at a time, test takers are presented with small groups of items, referred to as testlets. The CAT pool for this research consisted of 10 5-item testlets at each of three difficulty levels. The item parameters, which are statistics that describe the various item characteristics such as item difficulty, were specified to resemble those typically observed for the LAST. Using these item-level statistics, responses to the test questions were generated for simulated test takers. These simulations consisted of four conditions that varied in terms of group sample sizes and group ability distributions. Sample sizes for the two test taker groups were either 1,000 or 3,000. The distribution of test taker ability for the two groups was either the same or differed by one standard deviation. Results show the performance of the EB DIF approach to be very promising, even in extremely small samples. The EB estimates tended to be closer to their target values than did ordinary MH statistics; the EB statistics were also more highly correlated with the true DIF values than were the MH statistics. An appendix contains data tables. (Contains 9 figures, 3 tables, and 42 references.) (SLD)

LSAC RESEARCH REPORT SERIES

■ An Empirical Bayes Enhancement of Mantel-Haenszel DIF Analysis for Computer-Adaptive Tests

Rebecca Zwick
University of California, Santa Barbara

Dorothy T. Thayer
Educational Testing Service

■ Law School Admission Council
Computerized Testing Report 98-15
August 2003

A Publication of the Law School Admission Council

LSAC RESEARCH REPORT SERIES

An Empirical Bayes Enhancement of Mantel-Haenszel DIF Analysis for Computer-Adaptive Tests

Rebecca Zwick
University of California, Santa Barbara

Dorothy T. Thayer
Educational Testing Service

Law School Admission Council
Computerized Testing Report 98-15
August 2003

A Publication of the Law School Admission Council

The Law School Admission Council is a nonprofit corporation that provides services to the legal education community. Its members are 201 law schools in the United States and Canada.

This study is published and distributed by the Law School Admission Council (LSAC). The opinions and conclusions contained in these reports are those of the authors and do not necessarily reflect the position or policy of the Law School Admission Council.

Table of Contents

Executive Summary

The Law School Admission Council (LSAC) is currently investigating the feasibility of implementing a computer adaptive version of the Law School Admission Test (LSAT). The introduction of computer adaptive tests (CATs) requires that new approaches be developed for the analysis of item properties, including differential item functioning (DIF). DIF is said to occur when test takers from two demographic groups (say, men and women) perform differently on an item even after they have been matched in terms of overall ability. The presence of DIF may point to unintended sources of difficulty in the item (e.g., a math item may require sports knowledge that is more common in men than in women). A significant technical challenge in assessing DIF in CATs is the need to develop a method that will produce stable* results in small samples: Even if the total number of test takers for a CAT is large, the number of responses to some items may be very small.

Currently within the testing industry, the Mantel-Haenszel DIF analysis method is the most commonly used to detect DIF for paper-and-pencil tests. In fact, this is the analysis used for the LSAT. A body of statistical methods referred to as empirical Bayes (EB) methods are known to be capable of producing stable statistical results with fewer test takers. The present study investigated the applicability to CAT data of a DIF analysis method that involves an EB enhancement of the popular MH DIF analysis method.

The computerized LSAT test design assumed for this study was similar to that currently being evaluated for a potential computerized LSAT. Here, rather than being presented with a single test item at a time, test takers are presented with small groups of items, commonly referred to as testlets. The CAT pool for this research consisted of 10 five-item testlets at each of three difficulty levels. The item parameters, which are statistics that describe the various item characteristics such as item difficulty, were specified to resemble those typically observed for the LSAT. Using these item-level statistics, responses to the test questions were generated for simulated test takers. These simulations consisted of four conditions that vaired in terms of group sample sizes and group ability distributions; both of these factors are known to affect the performance of DIF methods. Sample sizes for the two test taker groups were either 1,000 or 3,000 (before application of the CAT algorithm). The distribution of test taker ability for the two groups were either the same or differed by one standard deviation.

The results showed the performance of the EB DIF approach to be very promising, even in extremely small samples. The EB estimates tended to be closer to their target values than did the ordinary Mantel-Haenszel (MH) statistics; the EB statistics were also more highly correlated with the true DIF values than were the MH statistics.

Introduction

Because computer adaptive tests (CATs) typically involve smaller samples than paper-and-pencil tests for at least some items, standard differential item functioning (DIF) techniques may not provide results with adequate precision in this setting. Zwick, Thayer, and Lewis (1997, 1999, 2000) developed an empirical Bayes (EB) approach to Mantel-Haenszel (MH) DIF analysis which yields more stable and interpretable results in small samples than do conventional procedures and, therefore, seems well suited to adaptive testing conditions. The computations involve the MH indexes and their standard errors, along with an assumed prior distribution for the true DIF parameters. In an earlier study, the EB methods were extensively investigated through simulation study and were applied to paper-and-pencil tests. The current report describes work that was conducted to investigate the applicability of these methods to a large-scale computer adaptive admission test. The study, which is sponsored by the Law School Admission Council (LSAC), is part of a research program (Pashley, 1997) that is intended to investigate the feasibility of a computer adaptive Law School Admission Test (LSAT).

The study, which addressed the technical innovations necessary for application of the EB DIF method to CATs, was based on a simulated test that involved a pool of adaptively administered five-item testlets. A scaled score was computed for each test taker based on responses to five of these testlets; this score was used as the matching variable for DIF analysis. Following this, the EB elaboration of the MH procedure was applied. The resulting DIF statistics were compared to the true (generating) DIF.

We are grateful to Law School Admission Council for their sponsorship, to Joyce Wang for assistance in data analysis, and to Charlie Lewis for consultation. Certain results of this study were presented at the annual meeting of the American Educational Research Association, San Diego, 1998. Some general information on the empirical Bayes method in this report, including tabular material, has appeared in slightly different form in publications by Zwick, Thayer, and Lewis (1997, 1999, 2000).

* An estimation procedure is considered stable if the estimates tend to be close to their target values.

This report addresses the role of Bayesian methods in psychometrics and educational research. It also describes the EB DIF model and presents some of the issues that had to be considered in applying the EB methods to CATs; describes the simulation study that was designed to investigate the CAT DIF procedures; presents the analyses and results; and finally, provides a discussion of the findings and outlines our ideas for future research.

Bayesian Methods in Psychometrics and Educational Research

Bayesian and empirical Bayes methods have found wide application in psychometrics and in educational research. For example, in studies of test validity, a Bayesian or EB approach can yield estimated regression coefficients that are more stable than are the usual least squares coefficients by pooling information from multiple schools. This pooling is achieved by making an assumption about the prior distribution of the true regression parameters across schools. To estimate the regression parameters for a particular school, this prior is combined with the school-level regression model. The mean of the resulting posterior distribution[1] is typically taken as the point estimate of the regression parameter for that school. (An analogous approach can be used to pool data from multiple years.) When the parameters of the prior are estimated from the data, a method of this kind is called *empirical* Bayes. In a fully Bayesian approach, a distribution would be assumed for the parameters of the prior.

A particularly lucid introduction to EB methods is given by Casella (1985); useful descriptions of EB philosophies and estimation methods are given by Efron and Morris (1973) and Braun (1989). EB regression models have been used in test validity studies, beginning with Rubin's important paper (1980) on the LSAT. Braun, Jones, Rubin, and Thayer (1983) discussed a general model for EB regression, which has been applied in several validity studies, such as Zwick (1993). An EB survival model developed by Braun and later modified (Braun & Zwick, 1993) was used to study time to Ph.D. candidacy and time to degree by Zwick and Braun (1988) and Zwick (1991). Charles Lewis and Dorothy Thayer have investigated EB methods for test equating. Finally, many of the hierarchical models that are now popular in educational research can also be characterized as Bayes or EB models. In EB DIF analysis, information is pooled across items to produce DIF estimates that are more stable than the original DIF statistics.

EB Enhancement of Mantel-Haenszel DIF Analysis

The Mantel-Haenszel DIF analysis procedure of Holland and Thayer (1988) is a well-established method for assessing DIF. A 2 x 2 x K table of test taker data is constructed based on item performance (right or wrong), group membership (the *focal group*, which is of primary interest, or the *reference group*), and score on an overall proficiency measure (with K levels). The MH (Mantel & Haenszel, 1959) odds ratio estimate is then used to compare the two groups in terms of their odds of answering the item correctly, conditional on the proficiency measure. The MH index of DIF, *MH D-DIF*, is obtained by multiplying the natural log of the MH odds ratio estimate $\hat{\alpha}_{MH}$ by -2.35; the transformation of $\hat{\alpha}_{MH}$ places *MH D-DIF* on the ETS delta scale of item difficulty (Holland & Thayer, 1985). By convention, *MH D-DIF* is defined so as to be negative when the item is more difficult for members of the focal group than it is for comparable members of the reference group. Phillips and Holland (1987) derived an estimated standard error for $\ln(\hat{\alpha}_{MH})$; their result proved to be identical to that of Robins, Breslow, and Greenland (1986).

The results of an MH DIF analysis typically include the *MH D-DIF* index, along with its estimated standard error. In making decisions about whether to discard items or flag them for review, however, testing companies may rely instead on categorical ratings of the severity of DIF. Several testing companies have adopted a system developed by ETS for categorizing the severity of DIF based on both the magnitude of the *MH D-DIF* index and the statistical significance of the results (see Zieky, 1993). According to this classification scheme, a "C" categorization, which represents moderate to large DIF, requires that the absolute value of *MH D-DIF* be at least 1.5 and be significantly greater than 1 (at $\alpha = .05$). A "B" categorization, which indicates slight to moderate DIF, requires that *MH D-DIF* be significantly different from zero (at $\alpha = .05$) and that the absolute value of *MH D-DIF* be at least 1, but not large enough to satisfy the requirements for a C item. Items that do not meet the requirements for either the B or the C categories are labeled "A" items, which are considered to have negligible DIF. Items that fall in the C category are subjected to further scrutiny and may be eliminated from tests. For most purposes, it is useful to distinguish between negative DIF (DIF against the focal group, by convention) and positive DIF (DIF against the reference group). This distinction yields a total of five DIF classifications: C-, B-, A, B+, and C+. We make use of this five-way categorization in our work, though we sometimes refer for convenience to the "A, B, and C categories."

[1] ayesian analysis, "posterior" means "following the collection of data."

Zwick, Thayer, and Lewis (1997, 1999, 2000) developed an empirical Bayes (EB) approach to DIF analysis and classification which yields more stable results in small samples than do conventional procedures and is therefore likely to be advantageous in adaptive testing conditions. An assumption is made about the prior distribution of DIF parameters across items. The prior is combined with the item's DIF results to estimate a posterior distribution; the posterior mean serves as the EB point estimate of the DIF parameter for that item.

In addition to offering an alternative point estimate of DIF, the EB method provides a version of the A, B, and C DIF classification system. Two related problems associated with the traditional classification approach are that (1) when sample sizes are small, the DIF category is unstable and may vary substantially from one test administration to another and (2) attaching an A, B, or C label to an item may convey the mistaken notion that an item's DIF category is deterministic. The EB approach yields an estimate of the *probability* that the true DIF for an item falls into the A, B, and C categories, based on an estimate of the posterior distribution of DIF parameters. The estimated A, B, and C probabilities can be regarded as representing our state of knowledge about the true DIF category for the item.

A possible advantage of the EB method of probabilistic DIF classification is that it may convey information about the sampling variability of DIF results in a more comprehensible way than do the current procedures. This alternative way of representing the variability of DIF findings lends itself well to graphical display. Pie charts can be used effectively to represent the posterior probabilities associated with the A, B, and C categories, as shown in the section, "Properties of EB Point Estimates." The EB methods can be modified easily if the current rules used to assign items to categories are adjusted or if other hypothesis-testing approaches are substituted for the Mantel-Haenszel procedure.

The EB approach to DIF analysis is related to three areas of previous research. A precursor to the method was developed in the context of a simulation study of DIF methods for computer adaptive tests conducted by Zwick, Thayer, and Wingersky (1994a, 1994b, 1995). Also, the variance component analysis of DIF developed by Longford and his colleagues (Longford, 1995, chapter 5; Longford, Holland, & Thayer, 1993) can be described as an EB approach. Finally, a Bayesian conceptualization of DIF was described by Holland in ETS internal documents (January 27, 1987; February 11, 1987).

Modification of the EB DIF methods for the LSAT CAT context required that we address several issues. First, we needed to take into account LSAC's interest in considering a CAT that was adaptive on the testlet level, rather than the item level; this required us to design a simulation that would involve testlet-based CAT administration. We then had to decide whether the matching of test takers for DIF analysis should be based on a score that took the testlet structure into account. Another determination we had to make was what set of items to use in estimating the prior distribution for the EB procedures, and whether our procedures for estimating the parameters of the prior, which had been developed for nonadaptive tests, needed modification for application to CATs. Finally, we had to determine whether the EB method, previously tested on samples no smaller than 200 test takers for the reference group and 50 for the focal group, could be applied successfully with even smaller samples. These issues are addressed in subsequent sections.

Statistical Model for the EB DIF Approach

The EB DIF method uses the observed values of *MH D-DIF* and *SE (MH D-DIF)*, along with an assumed prior distribution, to obtain the *posterior* distribution of true Mantel-Haenszel DIF parameters. The model can be expressed as follows. (The notation changes from *MH D-DIF* to *MH* and from *SE(MH D-DIF)* to *SE (MH)* to make the presentation less cumbersome.) We know that $\ln(\hat{\alpha}_{MH})$ has an asymptotic normal distribution (Agresti, 1990). Therefore, it is reasonable to assume that

$$MH_i|\theta_i \sim N(\theta_i, \sigma_i^2) \tag{1}$$

where MH_i is the MH statistic for item i, σ_i^2 is the sampling variance of the MH statistic, and $E(MH_i) = \theta_i$ represents the unknown parameter value corresponding to MH_i. In our computations, we assume that the sampling variance is known; that is, we set σ_i^2 equal to the observed estimate of the squared standard error, $SE^2(MH_i)$. The effect of ignoring the error associated with the estimation of $SE(MH_i)$ was judged to be minimal by Longford (1995); this was confirmed in analyses conducted by Zwick, Thayer, and Lewis (1997; 1999).

We assume the following prior distribution for θ_i:

$$\theta_i \sim N(\mu, \tau^2) \tag{2}$$

where μ is the across-item mean of the DIF parameters θ_i and τ^2 is the across-item variance. Estimation of μ and τ^2 is discussed in the section, "Estimation of μ and τ^2."

The posterior distribution of θ_i, given the observed statistic, MH_i, can be expressed as

$$f(\theta_i|MH_i) \propto f(MH_i|\theta_i)\, f(\theta_i) \tag{3}$$

Standard Bayesian calculations (see, e.g., Novick & Jackson, 1974) show that this distribution is normal with mean and variance given by

$$E(\theta_i|MH_i) = W_i MH_i + (1 - W_i)\mu \tag{4}$$

and

$$Var(\theta_i|MH_i) = W_i \sigma_i^2, \tag{5}$$

where

$$W_i = \frac{\tau^2}{\sigma_i^2 + \tau^2} \tag{6}$$

The means and variances for the model, prior, and posterior distributions are summarized in Table 1. The posterior mean is a *shrinkage* estimator of Mantel-Haenszel DIF, obtained by substituting estimates of μ and τ^2 for the corresponding parameters in equations 4–6 and setting σ_i^2 equal to $SE^2(MH_i)$. The larger the value of σ_i^2, the more the EB estimation procedure "shrinks" the observed MH value toward the prior mean (often zero or close to zero, as described in the section, Estimation of μ and τ^2). On the other hand, as σ_i^2 approaches zero, W_i approaches 1, and the posterior mean approaches the observed MH_i value.

TABLE 1
Means and variances of key distributions

Distribution	Mean	Variance
Model $f(MH_i\|\theta_i)$	θ_i (unknown)	σ_i^2 (treated as known and equal to $SE^2(MH_i)$)
Prior $f(\theta_i)$	μ	τ^2
Posterior $f(\theta_i\|MH_i)$	$W_i MH_i + (1 - W_i)\mu$	$W_i \sigma_i^2$

Note. $W_i = \frac{\tau^2}{\sigma_i^2 + \tau^2}$

Obtaining the posterior probabilities associated with the five possible true DIF categories (C-, B-, A, B+, and C+) is accomplished by considering a normal distribution with mean and variance equal to the estimates of the posterior mean and variance (equations 4 and 5), respectively. The magnitude criteria presented in the section, EB Enhancement of Mantel-Haenszel DIF Analysis, are then applied. For example, to estimate the probability that the true DIF category is C-, the area under this normal density function, which is to the left of -1.5, is obtained. (Since the goal of the procedure is to estimate the distribution of DIF *parameters*, the statistical significance criteria for C- status, described in that section are not relevant here.)

In summary, the steps in the EB DIF procedure are as follows:

1. For the item of interest, estimate the values of MH_i and $SE(MH_i)$. Assume the distribution of MH_i is normal, with unknown mean θ_i and known standard deviation $SE(MH_i)$.
2. Assume that the *prior distribution* for the true DIF parameter, θ_i, is normal. Use the observed mean and variance of *MH D-DIF* statistics across an appropriately defined set of items (a test section or form in the case of paper-and-pencil tests), along with an estimate of the across-item average of $SE^2(MH_i)$, to estimate the parameters of the prior (see Equation 7). The prior distribution is the same for the entire set of items.

3. Based on a and b, use standard Bayesian theory to estimate, for each item, the *posterior distribution* of the true DIF parameter θ_i, given the observed statistics. The posterior mean is the EB point estimate of DIF.
4. By applying the magnitude criteria associated with the DIF classifications to the posterior distribution, estimate the probabilities that the true DIF for the item is in each of the five DIF categories.[2]

Estimation of μ and τ^2

An aspect of the EB procedure that needs further explanation is the determination of reasonable values for the prior mean μ and the prior variance τ^2. When MH DIF analyses are conducted using number-right score as the matching variable, the MH DIF statistics are constrained to sum to approximately zero over the set of items. Therefore, in some applications, setting μ equal to zero may be appropriate. However, we have chosen to estimate μ as well as τ^2 from the current dataset; this less restrictive approach is appropriate under a wider range of circumstances, including analyses in which the matching variable is external to the test under investigation. Our estimates of μ and τ^2 were

$$\hat{\mu} = \text{Average}\left(MH_i\right) \quad \text{and} \quad \hat{\tau}^2 = \hat{Var}\left(MH_i\right) - \text{Average}\left(SE_i^2\left(MH_i\right)\right) \tag{7}$$

where $\hat{Var}\left(MH_i\right)$ is the observed across-item variance of the $\left(MH_i\right)$ statistics.

In our initial work, we justified these estimators as follows: Suppose that $MH_i = \theta_i + e_i$ as in the simplest version of the model of Longford, Holland, and Thayer (1993), where e_i is an error term with $E(e_i|\theta_i) = 0$ and $Var(e_i|\theta_i) = \sigma_i^2, i = 1, 2, \ldots n$. Suppose further that $Cov(e_i, e_j|\theta_i, \theta_j) = 0$ for $i \neq j$. Then $E(MH_i) = E(\theta_i) = \mu$ and $Var(MH_i) = \tau^2 + \sigma^2$, where $\sigma^2 = E(\sigma_i^2)$. These estimators can still be justified in the case of adaptive testing, where it may be unreasonable to assume that the σ_i^2 values have a common expectation. As described by Hoaglin, Mosteller, and Tukey (1991, p. 205), we can define σ^2 as $\sum_i \sigma_i^2 / n$ and estimate σ^2 as in the equal-variance case without losing much precision. We use the observed mean of the MH_i statistics as our estimate of μ. In estimating τ^2, we use Average $(SE^2(MH_i))$ to estimate σ^2 and use the observed across-item variance of the MH_i statistics as an estimate of $Var(MH_i)$. That is, in the formulation in Equation 7, the across-item prior variance of the true DIF values, τ^2 is estimated by deflating the estimated across-item variance of the DIF statistics by an amount equal to the average of the estimated item-level sampling variances. Similar estimators have been independently proposed by Camilli and his colleagues (e.g., Camilli & Penfield, 1997).

In the present study, we had to consider how to adapt our procedures for estimating μ and τ^2 for the CAT context. One determination we had to make was what set of items to use in estimating the prior distribution for the EB procedures; this is not entirely clear in the context of a testlet-based CAT. After considering several alternatives, we decided that the best way of preserving the advantages of the EB analysis without introducing unwieldy computational procedures was to estimate the parameters of the prior using data from all items in the pool.

In addition to deciding what set of items should be used in estimating the prior parameters, we needed to determine whether our former procedure for estimating τ^2 would perform well in the CAT context, where MH standard errors can vary considerably across items. In an unpublished simulation study, Dorothy Thayer and Charles Lewis compared our usual estimate of τ^2 (from Equation 7) to the iteratively obtained estimate of Longford (Longford, 1995; Longford, Holland, & Thayer, 1993) and to the approach developed by Camilli and his colleagues (see Camilli & Penfield, 1997). The simulation results showed that the seemingly unsophisticated estimate of Equation 7 performed best in a variety of circumstances. Estimation of μ and τ^2 in the current study is discussed further in the section, Comparison of $\hat{\mu}$ and $\hat{\tau}^2$ to Their Theoretical Values.

Validity Evidence

Zwick, Thayer, and Lewis (1997) conducted extensive validity studies of the EB DIF procedures, only a few of which are described here. Using simulated data, root mean square residuals (*RMSR*s) were computed to measure the deviation between DIF statistics and the true (generating) DIF, defined in the section, *True DIF* values and *True DIF* categories for Simulation Items. As expected, the EB point estimates had smaller *RMSR*s than did the ordinary *MH D-DIF* statistics; this advantage was greater in small samples. (See Casella, 1985, for a good intuitive explanation of the stability of EB estimates.) Application to actual test taker data for two administrations of the same test form showed that the Time 1 EB estimates were better predictors of the Time 2 MH_i statistics (i.e., had smaller *RMSR*s) than were the Time 1 MH_i statistics. Calibration plots

[2] Using computations only slightly more complex than those above, the EB approach can also be used to estimate the probability that an item will be classified as an A, B, or C in future administrations, based on the posterior predictive distribution (see Zwick, Thayer, & Lewis, 1997, 1999). The EB estimation procedures can also serve as the basis for a DIF detection procedure that uses loss functions (see Zwick, Thayer, & Lewis, 2000).

were used to examine the accuracy with which the EB procedure assigned items (probabilistically) to each DIF category. Similar plots are used to study the accuracy of weather predictions. Within every simulation condition in the study, one plot was constructed for each of the five DIF categories. The plot showed, for all items combined, the degree to which the EB-estimated probability of a particular DIF status (e.g., C-) corresponded to the true DIF status. Samples of these calibration plots appear in Zwick, Thayer, and Lewis (1997). Based on analysis of the plots, along with the *RMSR* evidence, we concluded that model fit was adequate. Analyses of the accuracy of the EB estimation procedures in the current study appear in the section, Analysis and Results.

Simulation Study

The simulation study was designed to capture properties of the populations and items involved in the LSAT. In addition, the item administration algorithm, which was based on adaptive testlet administration, was intended to be consistent with algorithms that are currently under consideration for the LSAT. The simulated CAT consisted of a pool of nonoverlapping five-item testlets. The pool contained 10 testlets at each of three difficulty levels, for a total of 30 testlets (comprising a total of 150 items). Five testlets were adaptively administered to each test taker, based on the testlet number-correct score, as described in the section, CAT Administration and Ability Estimation. This approach is similar to those currently under consideration by LSAC (Pashley, 1997; Schnipke & Reese, 1999). In particular, our design resembles that used in a simulation study by Reese, Schnipke, and Luebke (1999), which also involved the administration to each test taker of five 5-item testlets and included three testlet difficulty levels.

We initially considered using a scoring procedure that took the testlet structure into account. We ultimately decided, however, to use an item-response-theory based scale score for the entire item pool, as in Zwick, Thayer, and Wingersky (1994a, 1995), because this seemed most consistent with available LSAC scoring plans. The LOGIST computer program (Wingersky, Patrick, & Lord, 1988) was used to obtain an ability estimate for each test taker, based on the 25 items received by that test taker. The ability estimates were then transformed to scale scores expressed in the expected true score metric, as described in the section, CAT Administration and Ability Estimation. These scale scores were used for matching test takers for the DIF analysis. A summary of the simulation and analysis procedures appears in Table 2.

TABLE 2
Summary of simulation and analysis procedures

A. Create simulated test:
   1. Generate item parameters a, b, c, and $d = b_R - b_F$ for 150 items.
   2. Estimate a, b, and c parameters using LOGIST.
   3. Divide items into three groups of 50 based on estimated difficulty parameters.
   4. Within difficulty level, randomly assign the 50 items to 10 five-item testlets.

B. Generate all test takers to be used in the entire simulation: 600,000 test takers for the reference group and for each of the two focal groups (N(0,1) and N(-1,1)).

C. For each test taker in each simulation condition, do the following:
   1. Administer the 150-item test.
   2. Extract the responses to the 25 items that the test taker would have received under the CAT algorithm.
   3. Estimate abilities for the test taker based on the 25 responses and convert this estimate to the expected true score metric.

D. Match test takers on expected true scores and perform DIF analyses:
   1. Conduct MH DIF analyses for reference group compared to N(0,1) focal group (Condition 1) and reference group compared to N(-1,1) focal group (Condition 2). Use 3,000 test takers per group, which allows for 200 replications of each analysis. (3,000 x 200 = 600,000, the total number of test takers generated for each group.)
   2. Conduct another set of MH DIF analyses for reference group compared to N(0,1) focal group (Condition 3) and reference group compared to N(-1,1) focal group (Condition 4). This time, use 1,000 test takers per group and 600 replications of each analysis. (1,000 x 600 = 600,000.)
   3. Apply EB analyses to all MH results.

Simulation Conditions

The main portion of the simulation included four conditions, which differed in terms of focal group ability distribution (N(0, 1) or N(-1, 1))[3] and sample size per group (3,000 or 1,000). Each of these factors is detailed below; the four conditions are summarized in Table 3. To facilitate trouble-shooting and to allow pilot tests of our ability and DIF estimation procedures, we conducted a series of preliminary simulations, which are described in the section, Analysis of Data From Preliminary Simulations.

TABLE 3
Simulation conditions

Number	Focal Group Distribution	Initial *n* Per Group
1	N(0, 1)	3,000
2	N(-1, 1)	3,000
3	N(0, 1)	1,000
4	N(-1, 1)	1,000

Note. The notation N(x, y) refers to a normal distribution with mean x and y and variance y. In all cases, the reference group distribution was N(0, 1).

Ability Distributions

The reference group had a standard normal ability distribution in all simulation conditions. In Conditions 1 and 3, the focal group ability distribution was standard normal (N(0, 1)); in Conditions 2 and 4, it was N(-1, 1).

Sample Sizes

Under the CAT algorithm used here, the item sample sizes vary across testlets (although they are similar for testlets that share the same difficulty level). Pilot simulations were conducted to determine a value for the initial number of test takers that would produce a useful range of testlet sample sizes. One of our goals was to investigate the utility of the EB approach when group sample sizes fell below 100.

As noted earlier, we ultimately included two levels of sample size: one in which the initial sample size for each group was 3,000 (Conditions 1 and 2) and one in which the initial sample size per group was 1,000 (Conditions 3 and 4). The resulting item sample sizes are discussed in the section, Results of Main Simulation.

Model and Parameters for Data Generation

The data were generated using the three-parameter logistic (3PL) model. The probability of a correct response on item i in group G(G = R or F, denoting the reference or focal group) can be represented as

$$P_{iG}(\xi) = c_i + (1 - c_i)\{1 + \exp[-(1.7a_i(\xi - b_{iG}))]\}^{-1} \quad (8)$$

where ξ is the test taker ability parameter, a_i is the discrimination parameter for item i, C_i is the probability of correct response to item i for a very low-ability test taker (which is constant across items), and b_{iG} is the item difficulty in group G. The focal group difficulty, b_{iF}, is equal to $b_{iR} - d_i$. Hence, d_i is the difference between reference and focal group difficulties.

Ln(a) and b_R were assumed to have independent normal distributions across items: N(-.30, .10) and N(0, 1.25), respectively. The means and variances of these distributions were determined by examining the distributions of item parameter estimates based on two LSAT datasets. As in previous simulations (e.g., Zwick, Thayer, & Lewis, 1997; Zwick, Thayer, & Wingersky, 1994a), we used a fixed value of the guessing parameter, c, for all items. The value that was selected—.15—was approximately equal to the average estimated c value in the LSAT datasets. (The variances of the estimated c's in the two LSAT datasets were .0007 and .0112, respectively; therefore, the use of a constant value does not seem problematic.)

[3] The notation N(x, y) refers to a normal distribution with mean x and variance y.

We modeled d as N(0, .15), which produced reasonable results in terms of true A, B, and C status; see the following section. (LSAT DIF results were not available for use in modeling DIF.) The actual parameters for the 150 items used in the simulation (ordered by testlet number) appear in the Appendix for the three testlet difficulty levels; summary statistics (for the overall pool and for each testlet difficulty level) appear in Table 4.

TABLE 4
Descriptive statistics for the true parameters of the 150 simulation items

	Overall		Easy Testlets		Medium Testlets		Hard Testlets	
	Mean	Variance	Mean	Variance	Mean	Variance	Mean	Variance
ln(a)	-.28	.11	-.34	.10	-.19	.13	-.32	.10
b_R	.07	1.25	-1.12	.35	.05	.08	1.27	.44
d	.00	.14	-.06	.14	.06	.16	.01	.12
True DIF	.05	.93	-.19	1.08	.27	1.21	.06	.42

Note. The generating distributions for item parameters were N(-.30, .10) for ln(a), N(0, 1.25) for b_R, and N(0, .15) for d. All c parameters were set equal to .15. True DIF status for the items was as follows: 69.3% A items, 19.3% B items, and 11.3% C items.

True DIF Values and True DIF Categories for Simulation Items

As noted in Zwick, Thayer, and Lewis (1997), a simple relation between item parameters and MH DIF exists only in the Rasch model, in which the *MH D-DIF* statistic provides an estimate of $4ad$ under certain assumptions. In the present study, as in that study, we defined true DIF as follows:

$$True\,DIF = -2.35 \int \ln \left\{ \frac{P_{iR}(\xi)/Q_{iR}(\xi)}{P_{iF}(\xi)/Q_{iF}(\xi)} \right\} f_R(\xi)\, d\xi, \qquad (9)$$

where $P_{iG}(\xi)$ (G=R or F) is the item response function for group G, given by Equation 8, $Q_{iG}(\xi) = 1 - P_{iG}(\xi)$ and $f_R(\xi)$ is the reference group ability distribution. Pommerich, Spray, and Parshall (1995) proposed similar indexes in other contexts; Zwick, Thayer, and Mazzeo (1997) used a related definition in a study of DIF in polytomous items. Roughly speaking, this quantity can be viewed as the true MH value, unaffected by sampling or measurement error.[4] Empirical analyses, which included estimation of the regression of the mean Mantel-Haenszel statistics (across replications) on the *True DIF* values from Equation 9, supported this definition (Zwick, Thayer, & Lewis, 1997).

We defined an item's true DIF status in terms of its *True DIF* value, using only the magnitude criteria in the section EB Enhancement of Mantel-Haenszel DIF Analysis. That is, items with *True DIF* values exceeding 1.5 in magnitude were considered to have a true status of C, items with *True DIF* magnitudes between 1 and 1.5 were considered to have a true status of B, and the remaining items had a true status of A. The *True DIF* value and true DIF status for each of the 150 items in the simulation is given in the Appendix and summarized in Table 4. Of the 150 items, 69.3% were categorized as A items, 19.3% as B items, and 11.3% as C items.

Item Calibration and Development of Easy, Medium, and Hard Testlets

Though the generating item parameters are, of course, known in a simulation study, only estimates of these parameters are available in actual applications. To make the simulation more realistic, we obtained estimates of the item parameters to use in creating the E(asy), M(edium), and H(ard) testlets and in estimating test taker ability levels (see next section.) We calibrated the 150 items under no-DIF conditions, using 2,000 test takers from a standard normal ability distribution, with the LOGIST program (Wingersky,

[4] For conditions in which the reference and focal group had different ability distributions, we previously considered (Zwick, Thayer, & Lewis, 1997) an alternative definition of *True DIF* in which $f_R(\xi)$ was replaced by the mixture of reference and focal distributions, with mixing proportions determined by the relatve sample sizes. We decided against this index because (1) it seemed undesirable for *True DIF* to be defined so as to be dependent on the distributions and sample sizes for the two test taker groups and (2) substitution of the alternative definition would have made little difference in that study. The mixture-based function would be even more unwieldy in the CAT context, since the mixing proportions could potentially vary across items.

Patrick, & Lord, 1988), as in Zwick, Thayer, and Wingersky (1994a, 1994b, 1995). We then allocated the items to the E, M, and H testlets according to their estimated b_{iR} parameters. The items with the 50 lowest estimated b_{iR}'s were assigned to the E testlets, those with the 50 middle values were assigned to the M testlets, and those with the 50 highest values were assigned to the H testlets. Within each difficulty level, the 50 items were randomly assigned to 10 five-item testlets.

CAT Administration and Ability Estimation

Each test taker received a total of five testlets, starting with a randomly selected M testlet. The following rules governed which testlet the test taker received after the initial one:

1. If the number correct (on current testlet) is 0, 1, or 2, go down one step (or stay at E if already there)
2. If the number correct is 3, stay at current level
3. If the number correct is 4 or 5, go up one step (or stay at H if already there)
4. In each case, pick randomly from the testlets of the appropriate level, with the constraint that no testlet be given to an test taker more than once.

After the CAT had been administered, a maximum likelihood estimate (MLE) of test taker ability was estimated, using LOGIST, based on the 25 CAT items. (In the analyses of the preliminary dataset, described in the section Analysis of Data From Preliminary Simulations, an ability estimate was also obtained based on all 150 items for comparison purposes). These ability estimates were then converted to scale scores as follows:

$$\text{Scale score} = \sum_{i}^{150} \hat{P}_i(\hat{\xi}) \tag{10}$$

where $\hat{P}_i(\hat{\xi})$ is the estimated item response function for item i, evaluated at $\hat{\xi}$, the MLE of ability. This scale score is the expected true score for the entire pool of 150 items, as in Zwick, Thayer, and Wingersky (1994a; 1995). Scores of this type are also discussed by Stocking (1996). It was assumed that such scores would be used for reporting LSAT results if CAT administration were implemented.

Number of Replications Per Condition

The number of replications for each simulation condition depended on the sample size for that condition. For Conditions 1 and 2, for which the initial group sample sizes were 3,000, the number of replications was set to 200. For Conditions 3 and 4, for which the group sample sizes were 1,000, the number of replications was set to 600. This approach had the effect of roughly equalizing the standard errors of the across-replication means of the DIF statistics.

DIF Analysis

For each replication, the MH and EB DIF approaches were then applied. Because of the sparseness of tables, intervals of two units in the scale score metric were used in matching test takers. (Preliminary analyses showed that two-unit intervals produced more accurate DIF estimates than did one-unit intervals.) The mean and variance of the prior distribution were estimated using the MH results for all 150 items in the pool. For each item (in each simulation condition), the results consisted of EB point estimates of DIF, as well as the estimated probabilities of C-, B-, A, B+, and C+ status.

Analysis and Results

The description of analyses and results is organized into three main parts. First, the analysis of data from our preliminary simulations is presented. Then, the analysis of the properties of the EB point estimates, based on the main simulation, is described. Finally, the application of the probabilistic DIF classification analyses is detailed.

Analysis of Data From Preliminary Simulations

Preliminary simulations were conducted to check on the simulation procedures, the CAT algorithm, and the estimation routines; to investigate the relation between initial sample size and item sample size; and to conduct other analyses that would be impractical to apply to the data from the main simulation. The next two sections describe two main components of the preliminary analyses: an investigation of the properties ofthe CAT-based ability estimates and an examination of the performance of the EB DIF procedures in the no-DIF case.

Properties of Ability Estimates

To check on our ability estimation procedures, we first generated responses to all 150 items for 5,000 test takers from each of four populations: the reference group, which was standard normal (with no DIF, since the item difficulties for this group are the b_{iR} values), a standard normal focal group with DIF (as specified by the b_{iR} and d_i values given in the Appendix and summarized in Table 4), a N(-1, 1) focal group with no DIF, and a N(-1, 1) focal group with DIF.

The 25 items that each test taker would have received under the CAT algorithm were identified. LOGIST was then used to obtain, for each test taker, a CAT-based ability estimate, and, for comparison purposes, an ability estimate based on all 150 items. Residuals were obtained by subtracting the true ability values from the estimates. The distributions of the residuals were summarized using the median and the interquartile range (more appropriate than the mean and variance since LOGIST sets infinite or poorly determined ability estimates equal to arbitrary extreme values). Correlations between the ability estimates and true ability were also obtained. Results are given in Table 5, along with descriptive information about the true abilities.

TABLE 5
Properties of true and estimated abilities from preliminary simulation

	Reference N(0,1)	Focal N(0, 1) DIF present	Focal N(-1, 1) DIF absent	Focal N(-1, 1) DIF present
True ability				
Mean	-.012	-.007	-1.004	-1.010
SD	.987	1.006	1.005	1.001
Residual for 150-item estimate				
Median	.021	.027	.021	.015
IQR	.252	.267	.315	.327
Residual for CAT estimate				
Median	.033	.050	.024	.027
IQR	.522	.577	.615	.687
Correlations				
150-item with True ability	.979	.979	.959	.959
CAT with True ability	.919	.915	.888	.872
150-item with CAT	.940	.934	.919	.910

Note. The sample size was 5,000 for each of the 4 groups. *IQR* is interquartile range. Residuals are computed as *Ability Estimate - True Ability.*

The first two rows of the table show that the means and standard deviations of the true abilities were very close to their nominal values, supporting the accuracy of our data generation procedures. For the nonadaptive ability estimate, which was based on responses to all 150 items, the median residuals ranged from .015 to .027; in the CAT, the corresponding range was from .024 to .050. As expected, the median average residuals tended to be farther from zero for the CAT than for the 150-item test. In three of four comparisons, the departure from zero was greater when DIF was present than when it was not. The interquartile ranges (*IQR*s) show that the residuals were more variable for the CAT than for the 150-item test, more variable in the N(-1, 1) groups than the standard normal groups, and more variable when DIF was present. These findings are reasonable in light of the fact that short tests, DIF, and a poor match between group ability and test difficulty (as occurs in the N(-1, 1) groups) all tend to detract from measurement precision.

Correlations of the 150-item ability estimates with true abilities ranged from .96 to .98; for the CAT, these correlations ranged from .87 to .92. The intercorrelation between the two sets of estimated abilities ranged from .91 to .94. As in the case of the residuals, the obtained correlations were affected by the test length (confounded here with adaptive status), the location of the ability distribution, and to some degree, the presence of DIF.

Properties of EB DIF Estimates in the No-DIF Case

To check on our procedures and to determine the performance of the EB method in the null case, we created a data set in which DIF was absent. The abilities for both test taker groups were drawn from a standard normal distribution, and the reference group item difficulties were used in generating the data for both groups (i.e., the labeling of which group is the reference group and which is the focal group is entirely arbitrary). Both

a large-sample case, with 3,000 test takers per group and 200 replications, and a small-sample case, with 1,000 test takers per group and 600 replications, were considered. (Item sample sizes, of course, were smaller. The section, Results of Main Simulation, gives details on item sample sizes in the main simulation; results for the two groups in this preliminary simulation were similar to the reference group results in Table 7.)

Since the reference and focal groups had identical item response functions in this dataset, the *True DIF* value for each item was zero. How close were the EB DIF values to this target, and how did their accuracy compare to that of the ordinary MH statistics? We compared the EB point estimates of DIF to the standard MH statistics using root mean square residuals (*RMSR*s), defined for each item as follows:

$$RMSR=\sqrt{\frac{1}{Nrep}\sum_{j=1}^{Nrep}(Est\,DIF(j)-True\,DIF)^2}, \tag{11}$$

where *j* indexes replications, *Nrep* is the number of replications, *Est DIF(j)* is either the MH_i statistic or the EB posterior mean from the *j*th replication, and *True DIF* is the appropriate value from Equation 9; in this case, zero. The *RMSR* represents the average departure, in the MH metric, of the DIF estimate from the *True DIF* value. If these *True DIF* values are regarded as the estimands for the DIF statistics, then these *RMSR* values give estimates of the mean squared error (the average distance between the parameter estimate and the parameter) for the DIF statistics.

Table 6 summarizes the results for the 150 items in the two sample size conditions.[5] The table gives, for each condition, the 25th, 50th, and 75th percentiles of the distribution of *RMSR* values across the 150 items. The results are truly striking: The median *RMSR* for the MH method was roughly 10 times the median *RMSR* for the EB approach in both sample-size conditions. The EB DIF statistic departed from its target value of zero by an average of about .03 in the large-sample case and .07 in the small-sample case; the corresponding values for the MH were .37 and .68. The performance of the EB in the small-sample case was far superior to the performance of the MH in the large-sample case. Overall, the results for the no-DIF conditions suggest that the EB approach is useful for minimizing Type I errors in DIF analysis.

TABLE 6
Distribution of RMSRs across the 150 items for the no-DIF case from preliminary simulation

	Initial Group n = 1,000		Initial Group n = 3,000	
	EB	MH	EB	MH
25th percentile	.068	.543	.031	.298
Median	.072	.684	.034	.365
75th percentile	.078	.769	.037	.417

Note. The number of replications was 600 in the small-sample condition and 200 in the large-sample condition. Ability distributions for both groups were N(0, 1).

Results of Main Simulation

The main simulation explored the properties of the EB DIF point estimates and the probabilistic DIF classification procedures using data from the four simulation conditions summarized in Table 3. When the initial sample size was 3,000 (Conditions 1 and 2), item-level sample sizes (within a group) ranged from 86 to 842; for the initial sample size of 1,000 (Conditions 3 and 4), the range was from 16 to 307. Table 7 provides data on the realized sample sizes for each testlet difficulty level. (The CAT algorithm used in our study implies that item sample sizes within a difficulty level will be similar.) The table includes two test taker groups—the reference group, which has a standard normal distribution, and the focal group with a N(-1, 1) distribution. Sample sizes for the focal group with a standard normal distribution were similar to those of the reference group. In all cases, the effective sample size may be smaller, since the MH procedure automatically excludes cases for which no match is available in the other test taker group.

[5] In this unrealistic condition in which DIF was entirely absent and the true value of τ^2 was therefore equal to zero, our estimator of τ^2 took on a negative value in roughly half the replications, as is expected under this type of variance estimation procedure (Hoaglin, Mosteller, & Tukey, 1991, p. 210). We set these negative estimates to zero as is typically done when negative variance estimates are obtained in other contexts. We have never obtained a negative estimate of τ^2 in any of our analyses of actual test taker data.

TABLE 7
Testlet sample sizes for simulation study

	Large n		Small n	
	Focal N(0, 1): Condition 1	Focal N(-1, 1): Condition 2	Focal N(0, 1): Condition 3	Focal N(-1, 1): Condition 4
Easy testlets				
minimum	363	674	105	208
mean	424.9	765.1	141.6	255.0
maximum	494	842	191	307
Medium testlets				
minimum	653	544	194	160
mean	725.7	615.4	241.9	205.1
maximum	797	693	297	264
Hard testlets				
minimum	287	86	80	16
mean	349.4	119.5	116.5	39.8
maximum	406	158	151	67

Note. Each difficulty level contains 10 testlets. The results above are summaries over testlets and replications (600 replications in the small-sample condition; 200 replications in the large-sample condition). The reference group ability distribution was N(0, 1) in all conditions.

The smallest sample sizes occurred in the hard testlets, in which the mean item difficulty was quite high (1.27; see Table 4) relative to the abilities of the test takers, particularly for the focal group with a N(-1, 1) distribution.

Comparison of $\hat{\mu}$ and $\hat{\tau}^2$ to Their Theoretical Values

If the *True DIF* values of Equation 9 are regarded as the θ_i values, then the mean and variance of these quantities across items can reasonably be regarded as the target values for $\hat{\mu}$ and $\hat{\tau}^2$, respectively. The mean and variance across the 150 *True DIF* values (see Table 4) were 0.05 and 0.93, respectively. The distributions of $\hat{\mu}$ and $\hat{\tau}^2$ are summarized in Table 8 for the four simulation conditions. The median values of both $\hat{\mu}$ and $\hat{\tau}^2$ are somewhat higher than their target values, a finding that is not consistent with the results of our earlier EB DIF studies. (As noted in the section, Estimation of μ and τ^2, supplementary simulation work had shown that our procedure for estimating $\hat{\tau}^2$ produced better results than did competing approaches.) We plan to continue exploring ways to improve our procedures for estimating the parameters of the prior distribution. For example, a reviewer suggested that poorly determined MH_i statistics be trimmed before estimating τ^2, a procedure that might also improve the estimation of μ. Despite the apparent tendency to overestimate μ and τ^2, the EB point estimates performed well, as described in the subsequent sections.

TABLE 8
Distribution of $\hat{\mu}$ and $\hat{\tau}^2$ over replications

Simulation Condition			$\hat{\mu}$			$\hat{\tau}^2$		
Condition	Focal Group	Sample Size	25th Percentile	Median	75th Percentile	25th Percentile	Median	75th Percentile
1	N(0, 1)	Large	.07	.08	.09	1.08	1.12	1.17
2	N(-1, 1)	Large	.06	.08	.09	.94	.99	1.06
3	N(0, 1)	Small	.06	.08	.10	1.08	1.18	1.29
4	N(-1, 1)	Small	.05	.08	.11	.94	1.07	1.21

Note. The number of replications was 200 for Conditions 1 and 2 and 600 for Conditions 3 and 4. The mean and variance of *True DIF* values (Equation 9) are .05 and .93.

Properties of EB Point Estimates

Two analyses of the properties of the EB point estimates are reported for each of the four simulation conditions. The next section describes the computation of *RMSR*s for both EB and MH estimates; these indexes provide a measure of the average deviation of the DIF estimate from the *True DIF*. The section following that presents the correlations between the DIF estimates and *True DIF* values.

Comparison of RMSRs for MH D-DIF and for EB estimates of DIF. As in the preliminary simulation, we compared the EB point estimates of DIF to the MH_i statistics using root mean square residuals, defined in Equation 11. Table 9 summarizes the results for the 150 items in the four simulation conditions listed in Table 3. Table 9 gives, for each condition, the 25th, 50th, and 75th percentiles of the distribution of *RMSR* values across the 150 items, as well as the number of *RMSR* values exceeding one. In interpreting the *RMSR* values, it is useful to keep in mind that they are in the MH metric, and that the standard deviation of the *True DIF* values is .96.

TABLE 9
RMSR results for EB and MH DIF statistics in simulation study

	Large *n*				Small *n*			
	Focal N(0, 1): Condition 1		Focal N(-1, 1): Condition 2		Focal N(0, 1): Condition 3		Focal N(-1, 1): Condition 4	
	EB	MH	EB	MH	EB	MH	EB	MH
25th percentile	.284	.317	.302	.322	.460	.565	.464	.585
Median	.341	.390	.361	.366	.509	.713	.517	.641
75th percentile	.380	.444	.442	.594	.542	.787	.560	1.19
Number > 1	0	1	0	1	1	7	2	51

Note. Each *RMSR* summarizes results across replications. The results above are summaries over the 150 items. The reference group ability distribution was N(0, 1) in all conditions.

In the two small-*n* simulation conditions, the *RMSR* tended to be substantially smaller for the EB estimate than for *MH D-DIF*. The difference in median *RMSR* values was larger in the case in which both reference and focal ability distributions were standard normal. Here, the median *RMSR* for the EB was .51, compared to .71 for the MH. An even more striking finding was the difference in the number of *RMSR* values (out of 150) that exceeded one in the small-*n* conditions. When both reference and focal distributions were standard normal, the EB had one such *RMSR* value, compared to seven for the MH. When the focal group ability distribution was N(-1, 1), the EB had two *RMSR* values exceeding one, compared to 51 for the MH.

The small-*n* results were also examined separately for easy, medium, and hard items. As shown in Table 7, the smallest sample sizes occurred for the 50 hard items in Condition 4, when the focal group ability distribution was N(-1, 1). Here, reference group sample sizes ranged from 80 to 151 with a mean of 117; focal group sample sizes ranged from 16 to 67 with a mean of 40. These sample sizes are substantially smaller than is ordinarily considered acceptable for application of the MH procedure. Table 10 summarizes the *RMSR* results for these items. The median *RMSR* for the EB method for these items was .53, compared to 1.25 for the MH. The 25th and 75th percentiles for EB were .51 and .56; the corresponding values for the MH were 1.19 and 1.32. It is interesting to note that, in a subset of the results (not shown) for which the MH *RMSR* had a median of .53 (medium items, Condition 3), the sample sizes averaged about 240 per group. Roughly speaking then, the EB procedure achieved the same stability for samples averaging 117 and 40 reference and focal group members, respectively, as did the MH for samples averaging 240 per group. Table 10 also shows that, for the hard items in Condition 4, all 50 *RMSR*s for the MH procedure were greater than one, while only two of the 50 values exceeded one for the EB method.

TABLE 10
Distribution of RMSRs for the 50 hard items in condition 4 (small n, N(-1, 1) focal group)

	EB	MH
25th percentile	.514	1.190
Median	.532	1.252
75th percentile	.558	1.322
Number > 1	2	50

Note. The range of item sample sizes across the 50 items and 600 replications was from 80 to 151, with a mean of 117 for the reference group and from 16 to 67, with a mean of 40 for the focal group.

In the large-*n* conditions, the advantage of the EB estimates was greatly reduced. The lesser difference between the two DIF estimates is to be expected, since the MH standard errors are small when samples are large, causing the EB DIF estimate to be close to the MH values (see Equations 4 and 6).

The *RMSR* results are graphically displayed in Figures 1-3 for the easy, medium, and hard testlets, respectively. The plots show, for each of the 10 testlets at each difficulty level, the degree to which the median *RMSR* (across testlet items) for the MH exceeds the median *RMSR* for the EB method. The superiority of the EB approach is obvious in all three plots and is particularly notable in the hard testlets, especially in Condition 4 (as reflected in Table 10).

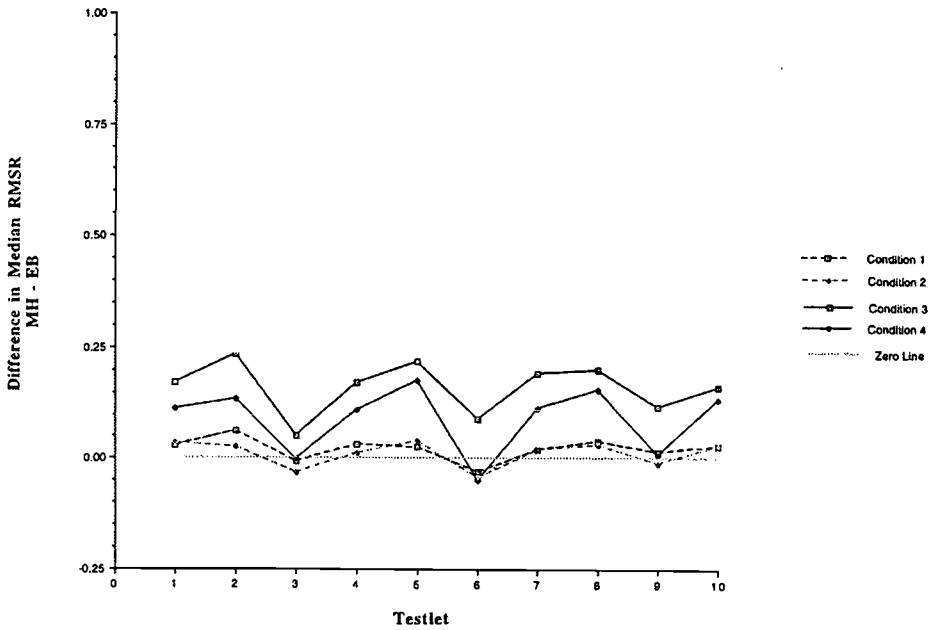

FIGURE 1. *Easy testlets—difference in median RMSR: MH-EB*

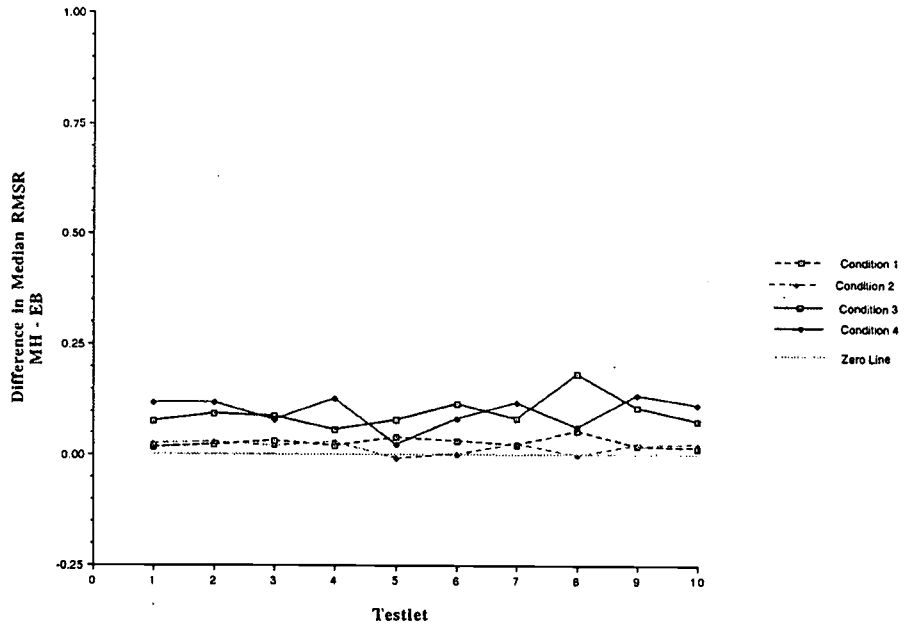

FIGURE 2. *Medium testlets—difference in median RMSR: MH-EB*

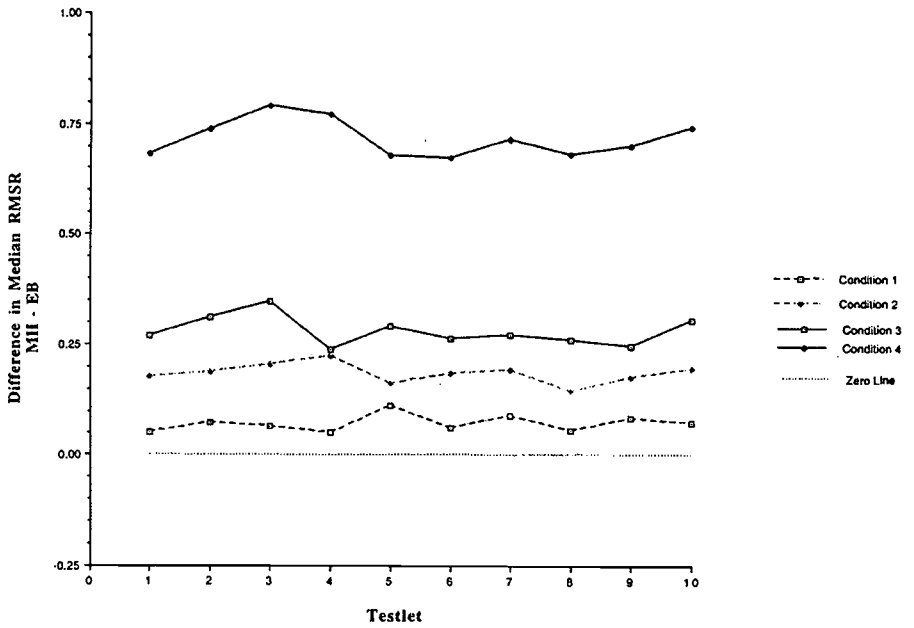

FIGURE 3. *Hard testlets—difference in median RMSR: MH-EB*

The generally smaller *RMSR* values for the EB estimates are consistent with theory. According to the *Stein effect* (named for statistician Charles Stein), "estimates can be improved by using information from all coordinates [in this case, the *MH D-DIF* values for all items] in estimating each coordinate" (Casella, 1985, p. 84). Such estimates have smaller mean squared error than their non-Bayesian counterparts. They are not, however, unbiased; in fact, the bias of these estimates is greatest for the extreme parameter values. The mean squared error of an estimator can be decomposed into two additive components: a variance term and a squared bias term. Our simulation design allows the estimation of each of these components. That is, the mean squared residuals (i.e., the squared *RMSR* values) can be decomposed into two additive terms: a variance term, estimated by the variability of the within-replication DIF estimates around their mean and a (squared) bias term, which is the squared difference between this mean and the *True DIF* value.

An analysis of the bias of the EB DIF statistics was conducted (as in Zwick, Thayer, & Lewis, 1997, 1999, 2000); results are displayed in Table 11. The 25th, 50th, and 75th percentiles of the bias values were similar for EB and MH in the large-sample conditions, but were larger for the EB method in the small-sample conditions, particularly when the reference and focal groups had different ability distribution. Surprisingly, in Conditions 1–3, the maximum MH bias was larger than the maximum EB bias; only in Condition 4 was the maximum bias value greater for the EB method.

TABLE 11
Bias and variance of DIF statistics for the 150 simulation items

	Large *n*				Small *n*			
	Focal N(0, 1): Condition 1		Focal N(-1, 1): Condition 2		Focal N(0, 1): Condition 3		Focal N(-1, 1): Condition 4	
	EB	MH	EB	MH	EB	MH	EB	MH
Variance								
Minimum	.050	.057	.065	.078	.144	.214	.155	.271
25th percentile	.079	.095	.084	.103	.195	.316	.191	.335
Median	.108	.141	.100	.127	.238	.498	.210	.402
75th percentile	.127	.166	.182	.339	.259	.592	.242	1.402
Maximum	.178	.352	.244	.575	.293	1.446	.296	2.347
Squared bias								
Minimum	.000	.000	.000	.000	.000	.000	.000	.000
25th percentile	.000	.001	.001	.000	.001	.001	.004	.000
Median	.001	.003	.003	.001	.007	.004	.027	.002
75th percentile	.019	.015	.016	.011	.035	.018	.088	.013
Maximum	.572	1.130	.325	.560	1.005	1.292	1.579	.697

Note. The variance and squared bias terms sum to the mean square residual (i.e., the squared *RMSR*). The reference group ability distribution was N(0, 1) in all conditions.

A possible concern about the EB method is that the extreme items tend to be most affected by the biasedness of the EB estimates. While this is a justifiable concern, the magnitude of the bias problem appears to be quite small in the present study. Several items for which EB bias was large still had smaller *RMSR* values for EB than for MH, showing that the EB statistics were, on the average, closer to their target values than the MH statistics. For example, the largest squared bias value for the EB approach, 1.58, occurred in Condition 4 for an item that had a *True DIF* value of 2.4. The EB variance value was .16, resulting in an *RMSR* value of 1.32. The MH statistic, by contrast, had a squared bias value of .69 and a variance term of 2.34, resulting in an *RMSR* value of 1.74. For only a very few items (6 in Condition 4, 2 in Condition 3, and none in Conditions 1 and 2) did the EB bias lead to EB *RMSR* values that exceeded MH *RMSR* values by more than 0.1.

Correlations of DIF estimates with True DIF values. As another check on the quality of the EB point estimates, the Pearson correlations between the EB estimates and the *True DIF* values were computed and compared to the correlations between the MH statistics and *True DIF*. The correlations between the two DIF estimates were also obtained. Correlations were computed for each replication within each simulation condition. Table 12 gives, for each condition the median correlation for the set of all replications within that condition. (Variability across replications was very small.) As expected, the correlations with *True DIF* were smaller for Conditions 3 and 4, the small-*n* conditions, than for Conditions 1 and 2. In each of the four conditions, the median correlation with *True DIF* was larger for the EB estimate than for the MH estimate. However, the differences were small in Conditions 1–3. In Condition 4, the small-*n* condition with the N(-1, 1) focal group, the EB estimate had a median correlation of .81 with *True DIF*, compared to only .75 for the MH statistic.

TABLE 12
Median correlations between DIF estimates and true DIF

	Large *n*		Small *n*	
	Focal N(0, 1): Condition 1	Focal N(-1, 1): Condition 2	Focal N(0, 1): Condition 3	Focal N(-1, 1): Condition 4
MH with *True DIF*	.932	.901	.830	.748
EB with *True DIF*	.934	.918	.837	.814
EB with MH	.999	.994	.991	.953

Note. The entries are the median correlations across replications. The number of replications was 200 for Conditions 1 and 2 and 600 for Conditions 3 and 4. The reference group ability distribution was N(0, 1) in all conditions.

Probabilistic DIF classification

Figure 4 gives a sample display of the results of a probabilistic DIF analysis. The plot is based on one replication for Item 138, which has a *True DIF* value of .43 and a true DIF status of A. The reference group sample size is 101; the focal group *n* is only 23. Given its MH value of 4.71, with a standard error of 2.22, this item would be classified as a C+ item using conventional rules. However, as the pie chart shows, the probabilistic approach produced the conclusion that there is about a 65% chance that the true status is A, a 20% chance that it is B+, and only a 14% chance that it is C+. (The estimated probabilities of B- and C- status sum to about 1%.) The EB point estimate of DIF for this item is .69, much closer to the *True DIF* value than is the MH estimate. This display illustrates the potential utility of the probabilistic DIF approach.

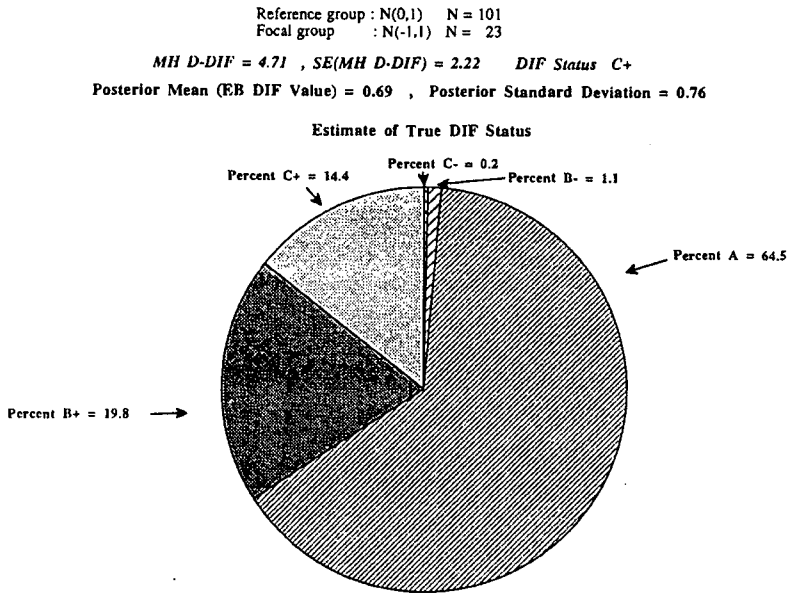

FIGURE 4. *DIF analysis of item 138 on replication 31 small sample size, true classification A*

To assess the accuracy of the probabilistic DIF classification methodology, we made use of calibration plots, which are illustrated in Figures 5–9. The calibration of prediction models was introduced in meteorological journals (e.g., Murphy & Epstein, 1967) and was subsequently addressed in the statistical literature (e.g., Dawid, 1982). The classic example is that of a weather forecaster who wishes to determine the accuracy of his predictions about the occurrence of rain. This can be done by considering the days on which the probability of rain was stated to be approximately equal to a certain value (say, 5%) and then determining, post hoc, on what percent of those days it did in fact rain. Ideally, that observed percent would be close to 5%. (See Gelman, Carlin, Stern, & Rubin, 1995, Chapter 6, for a description of similar types of model checking.)

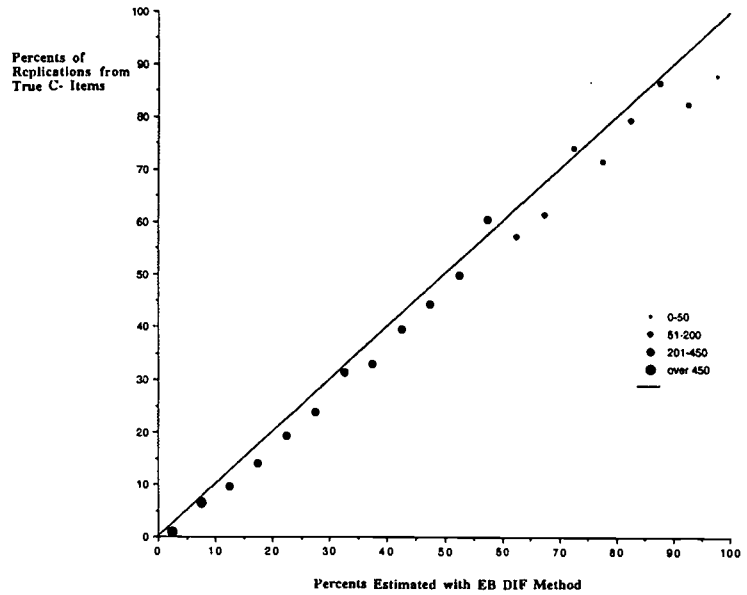

FIGURE 5. *Accuracy of EB probabilistic DIF classification for C- category, condition 4*

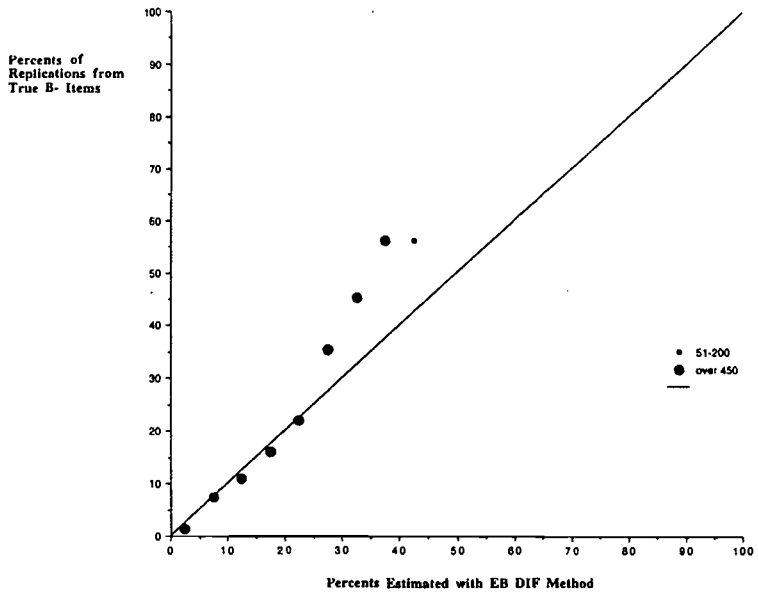

FIGURE 6. *Accuracy of EB probabilistic DIF classification for B- category, condition 4*

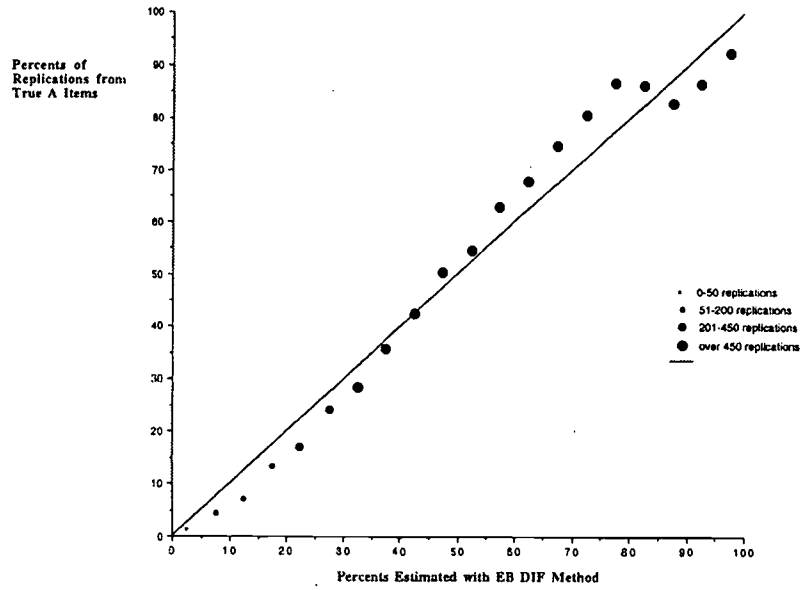

FIGURE 7. *Accuracy of EB probabilistic DIF classification for A category, condition 4*

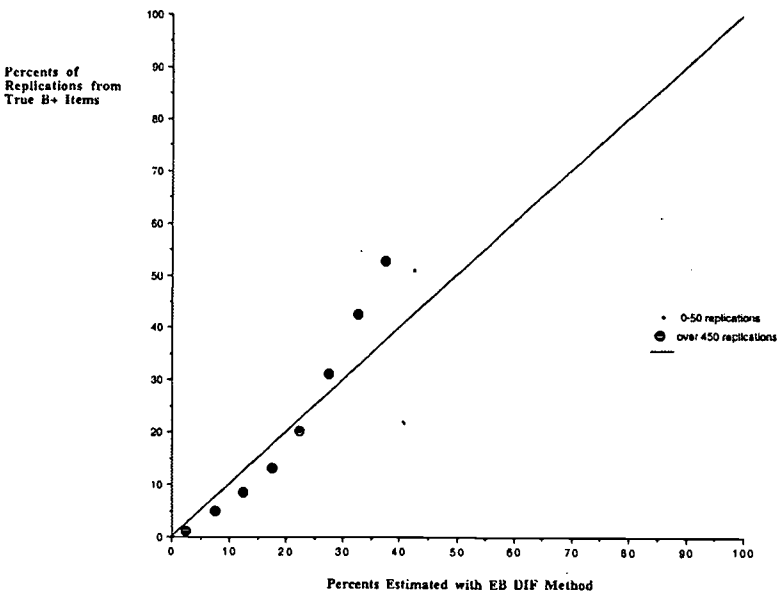

FIGURE 8. *Accuracy of EB probabilistic DIF classification for B+ category, condition 4*

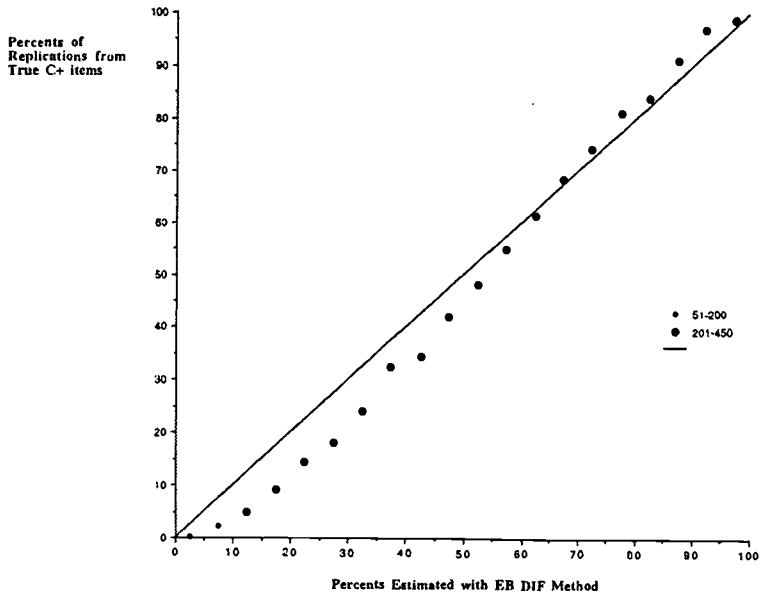

FIGURE 9. *Accuracy of EB probabilistic DIF classification for C+ category, condition 4*

In the present context, we wanted to examine the accuracy with which the EB procedure assigns items (probabilistically) to the A, B, and C DIF categories. We created a total of 20 calibration plots (4 simulation conditions x 5 DIF categories). For each plot, the data consist of the DIF results for the all replications of every item—a total of 200 x 150 = 30,000 observations in Conditions 1 and 2, and 600 x 150 = 90,000 observations in Conditions 3 and 4. Suppose we are interested in the C- category. Each replication can be classified according to the "percent C-" assigned using the EB DIF method; these values are plotted along the x-axis, grouped into intervals with a width of five percentage points. The values plotted along the y-axis correspond to the percentage of replications that come from an item with a true DIF status of C-. For example, suppose there are 1,000 replications for which the "percent C-" (based on the EB DIF method) is approximately 5%. If our method worked perfectly, we would find that 5% of these 1,000 replications (i.e., 50 of them) came from items with a true status of C-. (Obviously, this type of calibration can be applied only in simulations, where the true DIF status of the item is known.)

The five plots for Condition 4 are given in Figures 5–9 to illustrate the calibration technique. Condition 4 would be expected to produce the least favorable results because sample size is small and because the reference and focal group ability distributions are one standard deviation apart, a circumstance that is know to degrade the performance of DIF methods. Nevertheless, the results for the A category generally looked quite good; the "A" results for the other conditions were even better. (As the plots show, some points are poorly determined because they are based on a small number of replications.) The C- results also looked good; the C+ results, while still acceptable, did show some departures from the 45-degree line for x-values less than 60. Surprisingly, the C+ plots for the small-sample conditions (3 and 4) looked better than those for the large-sample conditions (1 and 2). The most disappointing results occurred in the B- and B+ plots. The patterns of results for the other three simulation conditions were quite similar to those obtained in Condition 4. We are investigating the reasons for the characteristic pattern in the B results, which did not occur in our earlier EB DIF research. Despite some anomalies, the 20 plots, considered as a group, led us to conclude that probabilistic DIF classification was working reasonably well. We are hopeful that modification of our methods for estimating μ and τ^2 and will lead to even better results for the probabilistic DIF classification approach.

Conclusions and Ideas for Future Research

The results of this study provided encouraging information about the stability of the empirical Bayes (EB) point estimates of DIF, even in very small samples. The EB estimates tended to be closer to their target values than did the ordinary Mantel-Haenszel (MH) statistics in terms of root mean square residual statistics (*RMSR*s);

the EB statistics were also more highly correlated with the *True DIF* values than were the MH statistics.

As theory would predict, the superiority of the EB approach was greatest in small samples. The smallest item sample sizes occurred for the 50 hardest simulation items when the initial group sample size was 1,000 and the reference and focal group ability distributions were one standard deviation apart. Here, item sample sizes for the reference group ranged from 80 to 151, with a mean of 117; focal group sample sizes ranged from 16 to 67 with a mean of 40. These sample sizes are substantially smaller than is ordinarily considered acceptable for application of the MH method. The *RMSR* results for these items showed that the EB statistics deviated from their target values by an average of .53 (in the Mantel-Haenszel DIF metric), compared to 1.25 for the MH. It is noteworthy that, in a different subset of the results, for which the MH *RMSR* had a median of .53, the item sample sizes averaged about 240 per group. Roughly speaking then, the EB procedure achieved the same stability for samples averaging 117 and 40 reference and focal group members, respectively, as did the MH for samples averaging 240 per group.

A possible drawback of the EB DIF method is that EB estimates are not unbiased; in fact, the bias of these estimates is greatest for the extreme parameter values. While this bias is a justifiable concern, analyses showed the magnitude of the bias problem to be fairly small in the present study. The 25th, 50th, and 75th percentiles of the distribution of bias values across items were similar for the EB and MH methods in the large-sample simulation conditions, but were larger for the EB method in the small-sample conditions, particularly when the reference and focal groups had different ability distributions. Surprisingly, however, in three of four simulation conditions, the maximum MH bias was larger than the maximum EB bias. Furthermore, some items with large EB bias still had smaller *RMSR* values for EB than for MH. For only a very few items (six and two, respectively, in the two small-sample conditions and none in the large-sample conditions) did the EB bias lead to EB *RMSR* values that exceeded MH *RMSR* values by more than 0.1.

Overall, the results suggest that it would be feasible to apply the EB DIF approach to adaptively administered LSAT items. As illustrated in this report, the EB point estimates of DIF can be supplemented with pie charts showing the probabilities associated with the A, B, and C categories of DIF severity.

We expect our future work to focus on the improvement of our estimation methods—particularly the procedures for estimating the variance parameter, τ^2. Although our current estimator for τ^2 performed better than its competitors in a supplementary simulation study, several avenues for improvement remain to be explored. For example, a reviewer suggested that, in estimating τ^2, MH values based on sparse data be excluded. We are considering the application of this idea, as well as other robust estimation procedures, to the estimation of both μ and τ^2. A possibly related goal for future research is an investigation of ways to refine the analysis procedures for the probabilistic DIF approach described in the section Probabilistic DIF Classification. Although the probabilistic DIF method performs adequately, there is clearly room for improvement.

The EB DIF method has been applied to the CAT version of the Armed Services Vocational Aptitude Battery (CAT-ASVAB) by Defense Department data analysts (Krass & Segall, 1998) and to simulation data by ACT researchers (Miller & Fan, 1998). Future applications by three other testing programs are under discussion. We expect to use the results of these applications to modify and refine the EB DIF analysis and estimation procedures. In addition, now that the testlet-based CAT simulation machinery has been created, it can be used to investigate the performance of the EB DIF methods under alternative CAT administration procedures.

References

Agresti, A. (1990). *Categorical data analysis*. New York: Wiley.

Braun, H. I. (1989). Empirical Bayes methods: A tool for exploratory analysis. In R. D. Bock (Ed.), *Multilevel analysis of educational data*. San Diego, CA: Academic Press.

Braun, H. I., Jones, D. H., Rubin, D. B., & Thayer. D. T. (1983). Estimation of coefficients in the general linear model from data of deficient rank. *Psychometrika, 48*, 171–181.

Braun, H., & Zwick, R. (1993). Empirical Bayes analysis of families of survival curves: Applications to the analysis of degree attainment. *Journal of Educational Statistics, 18*, 285–303.

Camilli, G., & Penfield, D. A. (1997). Variance estimation for differential test functioning based on Mantel-Haenszel statistics. *Journal of Educational Measurement, 34*, 123–139.

Casella, G. (1985). An introduction to empirical Bayes data analysis. *The American Statistician, 39*, 83–87.

Dawid, A. P. (1982). The well-calibrated Bayesian. *Journal of the American Statistical Association, 77*, 605–610.

Efron, B., & Morris, C. (1973). Combining possibly related estimation problems. *Journal of the Royal Statistical Society, Series B, 39*, 1–38.

Gelman, A., Carlin, J. B., Stern, H. S., & Rubin, D. B. (1995). *Bayesian data analysis*. London: Chapman & Hall.

Hoaglin, D. C., Mosteller, F., & Tukey, J. W. (1991). *Fundamentals of exploratory analysis of variance*. New York: Wiley.

Holland, P. W. (January 27, 1987). *Expansion and comments on Marco's rational approach to flagging items for DIF* (ETS internal memorandum). Princeton, NJ: Educational Testing Service.

Holland, P. W. (February 11, 1987). *More on rational approach item flagging* (ETS internal memorandum). Princeton, NJ: Education Testing Service.

Holland, P. W., & Thayer, D. T. (1985). *An alternative definition of the ETS delta scale of item difficulty* (ETS Research Report No. 85-43). Princeton, NJ: Educational Testing Service.

Holland, P. W., & Thayer, D. T. (1988). Differential item performance and the Mantel-Haenszel procedure. In H. Wainer and H. I. Braun (Eds.), *Test validity* (pp. 129–145). Hillsdale, NJ: Erlbaum.

Krass, I., & Segall, D. (1998). *Differential item functioning and on-line item calibration*. Draft report. Monterey, CA: Defense Manpower Data Center.

Longford, N. T. (1995). *Models for uncertainty in educational testing*. New York: Springer-Verlag.

Longford, N. T., Holland, P. W., & Thayer, D. T. (1993). Stability of the MH D-DIF statistics across populations. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning (p. 171–196)*. Hillsdale, NJ: Erlbaum.

Mantel, N., & Haenszel, W. (1959). Statistical aspects of the analysis of data from retrospective studies of disease. *Journal of the National Cancer Institute, 22*, 719–748.

Miller, T., & Fan, M. (April, 1998). *Assessing DIF in high dimensional CATs*. Paper presented at the annual meeting of the National Council on Measurement in Education, San Diego, CA.

Murphy, A. H., & Epstein, E. S. (1967). Verification of probabilistic predictions: A brief review. *Journal of Applied Meteorology, 6*, 748–755.

Nandakumar, R., & Roussos, L. (March, 1997). *Validation of CATSIB to investigate DIF of CAT data*. Paper presented at the annual meeting of the American Educational Research Association, Chicago.

Novick, M. R., & Jackson, P. H. (1974). *Statistical methods for educational and psychological research*. New York: McGraw-Hill.

Pashley, P. J. (1997). *Computerized LSAT research agenda: Spring 1997 update*. Newtown, PA: Law School Admission Council.

Phillips, A., & Holland, P. W. (1987). Estimation of the variance of the Mantel-Haenszel log-odds-ratio estimate. *Biometrics, 43*, 425–431.

Pommerich, M., Spray, J. A., & Parshall, C. G. (1995). *An analytical evaluation of two common-odds ratios as population indicators of DIF* (ACT Report 95-1). Iowa City, IA: American College Testing Program.

Reese, L. M., Schnipke, D. L., & Luebke, S. W. (1999). *Incorporating content constraints into a multistage adaptive testlet design* (Computerized Testing Report 97-02). Newtown, PA: Law School Admission Council.

Robins, J., Breslow, N., & Greenland, S. (1986). Estimators of the Mantel-Haenszel variance consistent in both sparse data and large-strata limiting models. *Biometrics, 42*, 311–323.

Rubin, D. B. (1980). Using empirical Bayes techniques in the law school validity studies. *Journal of the American Statistical Association, 75*, 801–816.

Schnipke, D. L., & Reese, L. M. (1997). *A comparison of testlet-based test designs for computerized adaptive testing*, (Computerized Testing Report 97-01). Newtown, PA: Law School Admission Council.

Stocking, M. (1996). An alternative method for scoring adaptive tests. *Journal of Educational and Behavioral Statistics, 21*, 365–389.

Wingersky, M. S., Patrick, R., & Lord, F. M. (1988). *LOGIST user's guide: LOGIST version 6.00* [Computer software]. Princeton, NJ: Educational Testing Service.

Zieky, M. (1993). Practical questions in the use of DIF statistics in test development. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 337–347). Hillsdale, NJ: Erlbaum.

Zwick, R. (1991). *Analysis of graduate school careers in three universities: Differences in attainment patterns across academic programs and demographic groups* (ETS Research Report 91-17). Princeton, NJ: Educational Testing Service.

Zwick, R. (1993). The validity of the GMAT for the prediction of grades in doctoral study in business and management: An empirical Bayes approach. *Journal of Educational Statistics, 18*, 91–107.

Zwick, R., & Braun, H. I. (1988). *Methods for analyzing the attainment of graduate school milestones: A case study* (ETS Research Report 88-30) Princeton, NJ: Educational Testing Service.

Zwick, R., Thayer, D. T., & Lewis, C. (1997) *An Investigation of the Validity of an Empirical Bayes Approach to Mantel-Haenszel DIF Analysis* (ETS Research Report No. 97-21). Princeton, NJ: Educational Testing Service.

Zwick, R., Thayer, D. T., & Lewis, C. (1999). An empirical Bayes approach to Mantel-Haenszel DIF analysis. *Journal of Educational Measurement, 36*, 1–28.

Zwick, R., Thayer, D. T., & Lewis, C. (2000). Using loss functions for DIF detection: An empirical Bayes approach. *Journal of Educational and Behavioral Statistics, 25*, 225–247.

Zwick, R., Thayer, D. T., & Mazzeo, J. (1997). Descriptive and inferential procedures for assessing DIF in polytomous items. *Applied Measurement in Education, 10*, 321–344.

Zwick, R., Thayer, D. T., & Wingersky, M. (1994a). A simulation study of methods for assessing differential item functioning in computerized adaptive tests. *Applied Psychological Measurement, 18*, 121–140.

Zwick, R., Thayer, D. T., & Wingersky, M. (1994b). *DIF analysis for pretest items in computer-adaptive testing* (ETS Research Report No. 94-33). Princeton, NJ: Educational Testing Service.

Zwick, R., Thayer, D. T., & Wingersky, M. (1995). Effect of Rasch calibration on ability and DIF estimation in computer-adaptive tests. *Journal of Educational Measurement, 32*, 341–363.

Appendix

TABLE A1
Parameters and parameter estimates for easy testlets

Item number	true a	true br	true d	estimated a	estimated br	estimated c	numer theo	class	TESTLET #
6	0.9432	-0.6478	0.5631	0.9951	-0.6332	0.1529	1.9566	C+	1
86	0.6279	-2.1791	-0.0464	0.6416	-2.1609	0.1529	-0.1137	A	1
87	0.8759	-1.6969	-0.6812	0.7955	-1.7835	0.1529	-2.2725	C-	1
95	0.8641	-0.787	0.359	0.9217	-0.7372	0.1529	1.148	B+	1
119	0.7134	-1.0456	0.1882	0.6514	-1.1015	0.1529	0.503	A	1
19	0.6827	-1.251	0.7938	0.6587	-1.2327	0.1529	2.0825	C+	2
56	0.8348	-1.8145	-0.0106	0.7965	-1.9324	0.1529	-0.0344	A	2
58	0.6146	-1.1669	0.4405	0.6407	-1.1054	0.1529	1.0254	B+	2
96	0.8107	-1.6927	-0.0001	0.7646	-1.7847	0.1529	-0.0003	A	2
123	1.1922	-0.3909	0.2825	1.3766	-0.3656	0.1574	1.167	B+	2
39	0.994	-0.9019	-0.2149	1.0124	-0.8156	0.1529	-0.7756	A	3
50	1.0053	-1.5142	-0.4295	1.0474	-1.5177	0.1529	-1.6426	C-	3
76	0.4188	-1.2148	-0.1883	0.4246	-1.0893	0.1529	-0.2912	A	3
78	0.7888	-0.8954	-0.4498	0.8273	-0.8138	0.1529	-1.2738	B-	3
132	0.8028	-0.6477	-0.4252	0.7995	-0.6508	0.1529	-1.1882	B-	3
42	0.6431	-0.7921	-0.542	0.6154	-0.7675	0.1529	-1.2346	B-	4
67	0.9064	-0.383	-0.5736	0.8794	-0.4457	0.1529	-1.6862	C-	4
89	0.7116	-1.0291	0.0264	0.6819	-1.1043	0.1529	0.0699	A	4
101	0.6297	-0.9507	0.4192	0.6127	-1.0478	0.1529	0.9875	A	4
148	0.4905	-1.702	-0.2092	0.5061	-1.6259	0.1529	-0.3896	A	4
33	0.7957	-1.334	-0.1196	0.8194	-1.285	0.1529	-0.3604	A	5
57	1.1962	-0.7447	0.0627	1.1854	-0.6574	0.1883	0.2705	A	5
75	0.6013	-1.152	0.4	0.5996	-1.104	0.1529	0.9088	A	5
131	0.477	-2.3123	0.4954	0.4865	-2.4049	0.1529	0.9215	A	5
142	0.6505	-1.6883	-0.174	0.6521	-1.7704	0.1529	-0.4339	A	5
9	0.6901	-1.982	-0.5036	0.6832	-1.9936	0.1529	-1.3412	B-	6
45	0.95	-0.3608	-0.4579	0.9786	-0.4067	0.0899	-1.4152	B-	6
91	1.0458	-1.878	-0.4751	1.0732	-1.744	0.1529	-1.9278	C-	6
94	0.6105	-0.6828	-0.8414	0.6057	-0.6271	0.1529	-1.7648	C-	6
107	0.7519	-2.2528	0.0354	0.7338	-2.2518	0.1529	0.1046	A	6
8	0.8295	-1.2986	-0.203	0.7792	-1.3514	0.1529	-0.6349	A	7
31	1.2054	-0.2784	-0.3637	1.1919	-0.367	0.1132	-1.3708	B-	7
92	0.7348	-2.3633	0.191	0.8845	-2.1087	0.1529	0.5534	A	7
130	1.1924	-0.5845	-0.0124	1.355	-0.4548	0.2575	-0.0517	A	7
134	0.7013	-1.3807	-0.067	0.6472	-1.3582	0.1529	-0.1783	A	7
4	0.8692	-0.7873	0.199	0.8441	-0.8223	0.1529	0.6355	A	8
51	0.6652	-0.9282	0.4908	0.7302	-0.8743	0.1529	1.2236	B+	8
136	0.5713	-0.4367	-0.0891	0.5686	-0.4013	0.1529	-0.1784	A	8
140	0.3841	-1.0101	-0.0178	0.3619	-0.9821	0.1529	-0.025	A	8
149	0.7588	-1.1214	-0.0176	0.7863	-1.2107	0.1529	-0.05	A	8
13	1.2218	-2.0929	-0.4201	1.2233	-2.2343	0.1529	-2.0146	C-	9
22	1.0245	-1.4424	0.1774	0.9778	-1.4842	0.1529	0.7016	A	9
24	0.5265	-1.3438	0.2043	0.5066	-1.3374	0.1529	0.4067	A	9
81	0.5514	-0.43	-0.7218	0.5096	-0.411	0.1529	-1.3418	B-	9
117	0.6713	-0.6045	-0.5035	0.7339	-0.6025	0.1529	-1.1718	B-	9
12	0.4083	-0.7616	0.219	0.4059	-0.7307	0.1529	0.3254	A	10
29	0.5071	-0.2553	0.063	0.4036	-0.4411	0.1529	0.111	A	10
54	0.5345	-0.3707	0.3029	0.4549	-0.3837	0.1529	0.5752	A	10
60	0.3606	-0.5874	-0.1286	0.3333	-0.6742	0.1529	-0.1648	A	10
71	0.4313	-1.0404	0.0721	0.4026	-1.1503	0.1529	0.1147	A	10

TABLE A2
Parameters and parameter estimates for medium testlets

item number	true a	true br	true d	estimated a	estimated br	estimated c	numer theo	class	testlet #
52	0.3295	-0.3978	0.5873	0.3308	-0.3439	0.1529	0.694	A	1
20	0.9763	-0.2277	-0.3668	1.0039	-0.3163	0.1203	-1.1381	B-	1
77	0.6428	-0.0355	0.1997	0.615	-0.009	0.1529	0.4331	A	1
69	0.9311	-0.0702	-0.0564	1.0595	0.043	0.2008	-0.168	A	1
59	0.7793	0.2706	-0.0465	0.8961	0.2952	0.1594	-0.1099	A	1
115	0.9769	-0.0839	0.132	0.8896	-0.1917	0.0874	0.4203	A	2
11	0.5841	-0.2964	0.2577	0.5981	-0.1913	0.1529	0.529	A	2
66	0.8289	-0.3901	0.2698	1.0262	-0.0661	0.2982	0.7878	A	2
18	0.703	0.2497	-0.3381	0.6502	0.1894	0.1529	-0.7131	A	2
73	0.4435	0.3661	0.1262	0.37	0.4449	0.1529	0.1831	A	2
5	0.9413	-0.1419	0.3238	0.89	-0.2023	0.0952	1.0286	B+	3
114	0.5936	-0.2792	-0.5829	0.5735	-0.187	0.1529	-1.1477	B-	3
41	0.8343	-0.1494	0.1493	0.9003	-0.0682	0.1604	0.4184	A	3
36	0.8662	0.2095	0.411	0.9584	0.3399	0.1859	1.1368	B+	3
93	1.0855	0.5871	-0.0772	0.9865	0.4506	0.0884	-0.2065	A	3
80	0.5406	-0.0191	0.738	0.5306	0.0862	0.1529	1.3972	B+	4
144	0.7417	0.086	0.1619	0.7685	0.1018	0.1214	0.3908	A	4
141	1.4698	0.4929	0.0396	1.633	0.4959	0.1465	0.138	A	4
122	1.0248	0.3835	0.7852	1.2595	0.5073	0.197	2.5093	C+	4
125	0.8785	0.4518	0.0045	0.9398	0.5344	0.1731	0.0112	A	4
99	0.8535	-0.4054	-0.4591	1.0353	-0.1811	0.2423	-1.3006	B-	5
79	0.7993	-0.0808	-0.7611	0.8485	-0.0621	0.1758	-1.8284	C-	5
32	0.5297	-0.0382	0.5672	0.4518	-0.0547	0.1529	1.046	B+	5
121	1.1298	0.0673	-0.3114	1.3043	0.2411	0.1957	-1.0063	B-	5
104	0.6435	0.1888	-0.6064	0.7423	0.3552	0.2165	-1.1707	B-	5
88	1.1042	-0.1578	0.7504	1.0589	-0.1995	0.1246	2.8759	C+	6
7	0.8502	-0.2773	-0.1033	0.9546	-0.1948	0.1983	-0.2946	A	6
103	0.4724	-0.0795	0.1157	0.4325	-0.0023	0.1529	0.1874	A	6
49	0.8738	0.1718	-0.4868	0.8426	0.1432	0.1117	-1.2148	B-	6
83	1.2568	0.4927	-0.2019	1.0806	0.4907	0.1197	-0.5994	A	6
26	1.2642	-0.3638	0.0663	1.2907	-0.3484	0.1498	0.2815	A	7
16	0.646	-0.0527	-0.0973	0.6279	-0.0294	0.1529	-0.2078	A	7
28	0.6131	0.1384	-0.0558	0.6425	0.0763	0.1529	-0.1105	A	7
21	0.8277	0.1466	0.5112	0.8285	0.1972	0.1607	1.3901	B+	7
147	0.7482	0.4239	-0.7829	0.7417	0.4646	0.1638	-1.5393	C-	7
113	0.993	-0.3734	0.4165	0.9744	-0.2733	0.1625	1.4579	B+	8
135	1.0399	-0.1754	0.6361	0.9927	-0.2264	0.1215	2.2904	C+	8
84	1.1017	0.2681	-0.3372	1.0846	0.2261	0.1025	-0.987	A	8
25	1.4613	0.2623	0.0769	1.4257	0.2729	0.1284	0.3015	A	8
34	1.4901	0.4412	0.379	1.5596	0.4511	0.1325	1.4873	B+	8
116	2.1779	-0.3198	0.3066	2	-0.3417	0.1223	2.169	C+	9
17	0.5619	-0.1782	-0.3579	0.5675	-0.1677	0.1529	-0.6715	A	9
145	1.1593	-0.0126	0.2744	1.1583	0.0176	0.1462	1.0152	B+	9
47	1.1481	0.1298	0.2029	1.058	0.2011	0.1655	0.7083	A	9
40	0.9801	0.564	0.059	0.9205	0.5678	0.1243	0.1535	A	9
139	0.7989	0.0149	-0.2377	0.704	-0.0526	0.1529	-0.5958	A	10
23	0.7046	-0.0929	0.345	0.829	-0.0392	0.179	0.8311	A	10
30	0.4712	0.1067	-0.1424	0.4226	0.1253	0.1529	-0.2218	A	10
106	0.9364	0.2118	0.7709	1.0756	0.2998	0.2115	2.3682	C+	10
90	0.4513	0.4675	-0.0637	0.5223	0.5156	0.1529	-0.0914	A	10

TABLE A3
Parameters and parameter estimates for hard testlets

Item number	true a	true br	true d	estimated a	estimated br	estimated c	numer theo	class	testlet #
108	0.7608	0.4004	0.0764	0.8321	0.582	0.2113	0.1741	A	1
64	0.656	0.6412	-0.0849	0.6132	0.7102	0.1529	-0.159	A	1
150	0.7828	0.6618	-0.1678	0.8476	0.7901	0.1746	-0.3485	A	1
72	0.7973	0.8877	-0.3661	0.7872	0.9106	0.1275	-0.6791	A	1
111	0.3454	1.9564	-0.0322	0.355	2.0195	0.1529	-0.0296	A	1
100	1.057	0.5118	-0.0898	1.1881	0.6224	0.1975	-0.2434	A	2
63	0.9154	0.7982	0.1511	1.2344	0.8352	0.1666	0.3494	A	2
65	0.5156	1.1304	-0.0546	0.5871	1.3404	0.1853	-0.0767	A	2
27	0.861	2.0069	-0.4846	1.1049	2.0401	0.1843	-0.447	A	2
85	0.7161	2.0542	0.4642	0.9135	2.1676	0.197	0.583	A	2
68	0.904	0.631	0.1414	0.9599	0.637	0.154	0.3449	A	3
37	0.7951	1.097	-0.0911	0.7586	1.1133	0.1377	-0.1637	A	3
98	0.6609	2.6982	0.4937	0.6205	2.409	0.1158	0.438	A	3
15	0.6651	2.8122	0.2804	0.5981	3.2314	0.1444	0.2158	A	3
143	0.8288	2.939	-0.1909	0.4647	3.629	0.1236	-0.0905	A	3
3	0.9596	0.7093	-0.0969	0.8774	0.71	0.1493	-0.2271	A	4
137	0.8378	1.0615	0.5225	0.916	1.0995	0.1362	1.1094	B+	4
10	0.774	1.0737	0.0878	0.6384	1.1938	0.155	0.1632	A	4
127	0.368	1.2141	0.0761	0.6866	1.6034	0.2994	0.0834	A	4
92	0.8466	2.1516	0.012	0.9523	2.2096	0.1508	0.012	A	4
128	0.7817	0.9651	-0.3593	0.7785	0.8185	0.102	-0.6399	A	5
110	0.8756	0.8802	0.9605	0.8711	0.8257	0.1248	2.4035	C+	5
53	0.7159	1.4123	-0.1146	0.6253	1.3056	0.1094	-0.171	A	5
74	0.726	2.1769	0.0661	0.7171	2.1504	0.1452	0.069	A	5
55	0.6071	2.2799	-0.6038	0.4305	2.6696	0.1231	-0.5103	A	5
138	0.5926	0.7947	0.2456	0.5213	0.5929	0.0689	0.4273	A	6
97	0.4772	0.6487	0.0305	0.4386	0.6378	0.1529	0.045	A	6
146	1.0007	1.0448	0.3219	1.1407	1.0443	0.1656	0.7278	A	6
129	0.4451	1.4136	-0.4984	0.4142	1.655	0.1529	-0.5661	A	6
102	1.3882	1.7001	-0.0768	1.286	1.8307	0.1535	-0.0965	A	6
120	0.7069	0.4557	0.9464	0.7353	0.5687	0.1659	2.1895	C+	7
70	0.5047	0.8107	0.3793	0.4204	0.9719	0.1529	0.5867	A	7
1	0.4785	1.0151	-0.2621	0.4491	1.1039	0.1529	-0.351	A	7
43	1.584	1.3512	0.3798	1.6542	1.4092	0.1508	0.8455	A	7
14	0.6469	1.3696	-0.3507	0.5906	1.5164	0.1453	-0.4898	A	7
118	0.5175	0.8228	-0.1571	0.5595	0.8798	0.1529	-0.2347	A	8
35	1.0353	0.8614	-0.2288	1.1757	0.8944	0.1736	-0.5014	A	8
124	0.8031	0.9034	-0.2354	0.9872	1.0148	0.1846	-0.4475	A	8
62	0.5841	0.918	-0.2285	0.6826	1.1376	0.1952	-0.3597	A	8
44	0.4555	1.1998	-0.383	0.5579	1.6123	0.2209	-0.4702	A	8
126	0.664	1.0294	-0.0342	0.6426	1.0026	0.1313	-0.0581	A	9
109	0.8492	1.1652	-0.3038	0.8996	1.0983	0.1551	-0.5152	A	9
133	0.718	1.0862	0.3229	0.7082	1.3307	0.1758	0.5976	A	9
105	1.0927	1.5521	0.5524	1.6201	1.5064	0.1617	1.029	B+	9
46	0.7271	1.7441	-0.1353	0.7913	1.8058	0.1875	-0.171	A	9
61	0.9598	0.7243	0.0145	1.0108	0.7992	0.1827	0.0346	A	10
38	0.7324	0.6478	-0.5742	0.7481	0.8117	0.1807	-1.0709	B-	10
112	0.6218	1.0758	-0.0125	0.6951	1.1304	0.1599	-0.0202	A	10
2	1.1578	1.2702	-0.1525	1.2583	1.3494	0.1618	-0.273	A	10
48	0.5315	2.8936	0.219	0.8348	2.4936	0.1816	0.1728	A	10

U.S. Department of Education
Office of Educational Research and Improvement (OERI)
National Library of Education (NLE)
Educational Resources Information Center (ERIC)

NOTICE

Reproduction Basis

[X] This document is covered by a signed "Reproduction Release (Blanket)" form (on file within the ERIC system), encompassing all or classes of documents from its source organization and, therefore, does not require a "Specific Document" Release form.

[ ] This document is Federally-funded, or carries its own permission to reproduce, or is otherwise in the public domain and, therefore, may be reproduced by ERIC without a signed Reproduction Release form (either "Specific Document" or "Blanket").